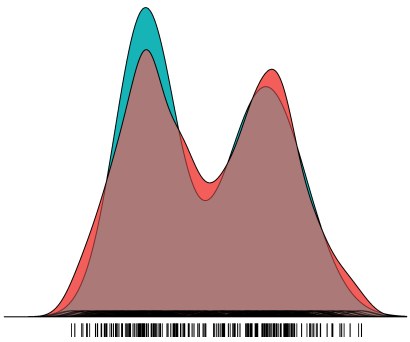
TEMA 2: Estimadores del núcleo de la función de densidad



José R. Berrendero

Departamento de Matemáticas, Universidad Autónoma de Madrid

Temas a tratar

- Definición de los estimadores del núcleo
- Representación gráfica y cálculo
- Interpretación como una convolución
- Error cuadrático medio integrado
- Selección del parámetro de suavizado
- Estimación de densidades multivariantes

Objetivo

- $X_1, \ldots, X_n$ iid según distribución F con densidad f
- El objetivo es estimar f sin hacer hipótesis previas sobre ella
- Dejar que los datos *hablen por sí mismos*
- Aparecen algunas nociones recurrentes en estadística:
  - **Sesgo y varianza** y la importancia de equilibrarlos
  - **Necesidad de seleccionar una constante** (*tuning constant* o *regularization parameter*) de la que dependen crucialmente las propiedades
  - La **maldición de la dimensionalidad** que puede ocurrir con datos de alta dimensión

Estimadores del núcleo

El objetivo es calcular un estimador $\hat{f}$ a partir de los datos tal que $\hat{f} \approx f$

Si $h \approx 0$,

$$\mathrm{P}(x-h \leq X \leq x+h) = \int_{x-h}^{x+h} f(t)dt \approx 2hf(x)$$

Es decir,

$$f(x) \approx \frac{1}{2h}\mathrm{P}(x-h \leq X \leq x+h)$$

Sustituyendo la probabilidad por la correspondiente proporción se obtiene un estimador de la densidad

Estimadores del núcleo

El resultado es:

$$\hat{f}\,(x) = \frac{1}{2h}\frac{\#\{i:\,|x-X_i|<h\}}{n} = \frac{1}{2hn}\#\left\{i:\,\frac{|x-X_i|}{h}\right.$$

Si $K(x) = (1/2)\mathbb{I}_{\{|x|\leq 1\}}$, donde $\mathbb{I}_A$ denota la función indicatriz sobre A,

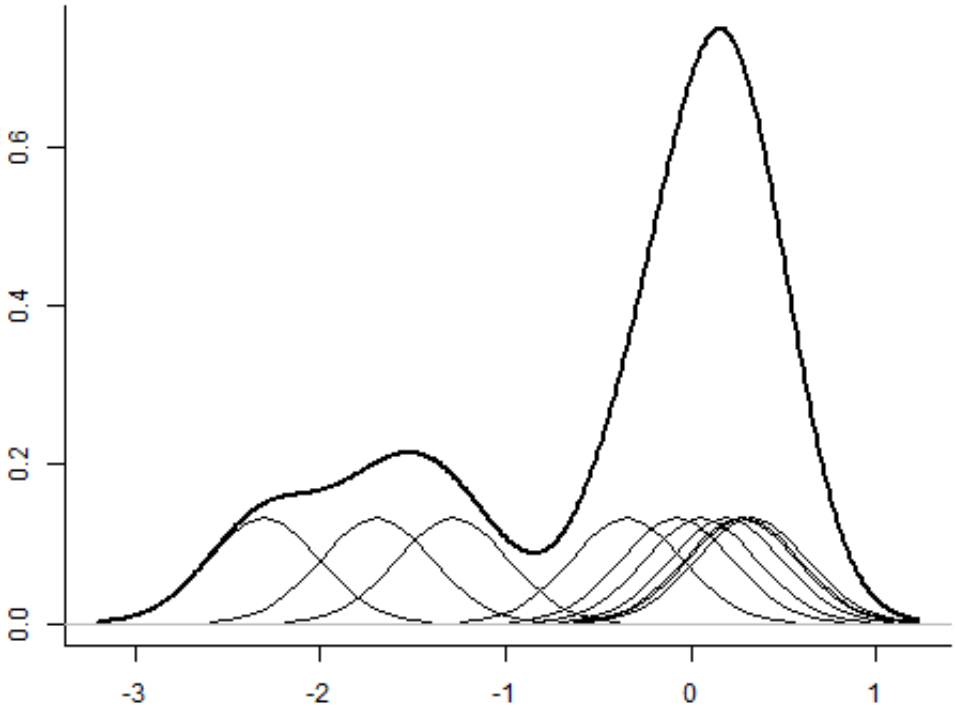
$$\hat{f}\,(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right).$$

Sustituyendo la indicatriz por otra función más suave obtendremos aproximaciones que se asemejan más a una densidad típica:

$$\hat{f}\,(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$$

- $h>0$ es el **parámetro de suavizado**
- K es el núcleo (que suele ser una función de densidad: $K\geq 0$ y $\int K = 1$)

Estimadores del núcleo



Núcleo gaussiano con $h = 0.3$

Representación con `geom_density`

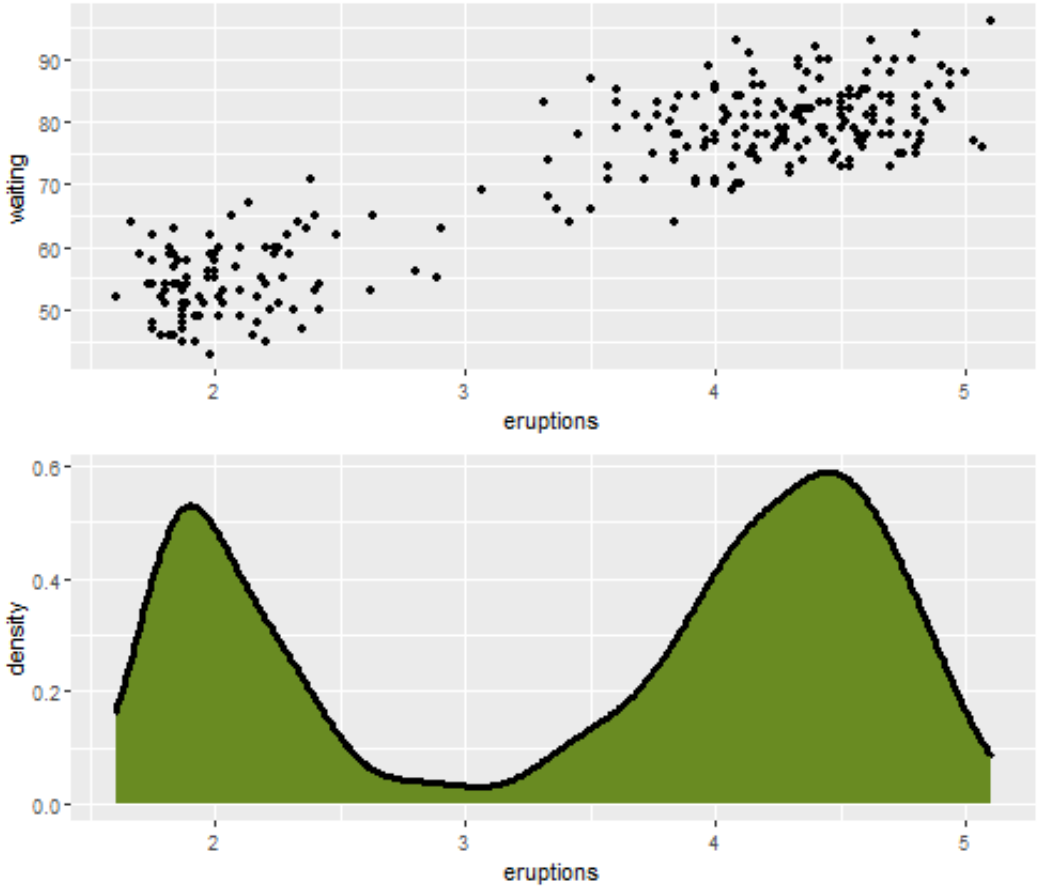
Fichero `faithful`: duración de las erupciones y del tiempo hasta la siguiente erupción del geyser *Old Faithful*

- `bw = 0.15` fija el valor de h.
- el núcleo por defecto es gaussiano (para cambiarlo, se usa el argumento `kernel`)

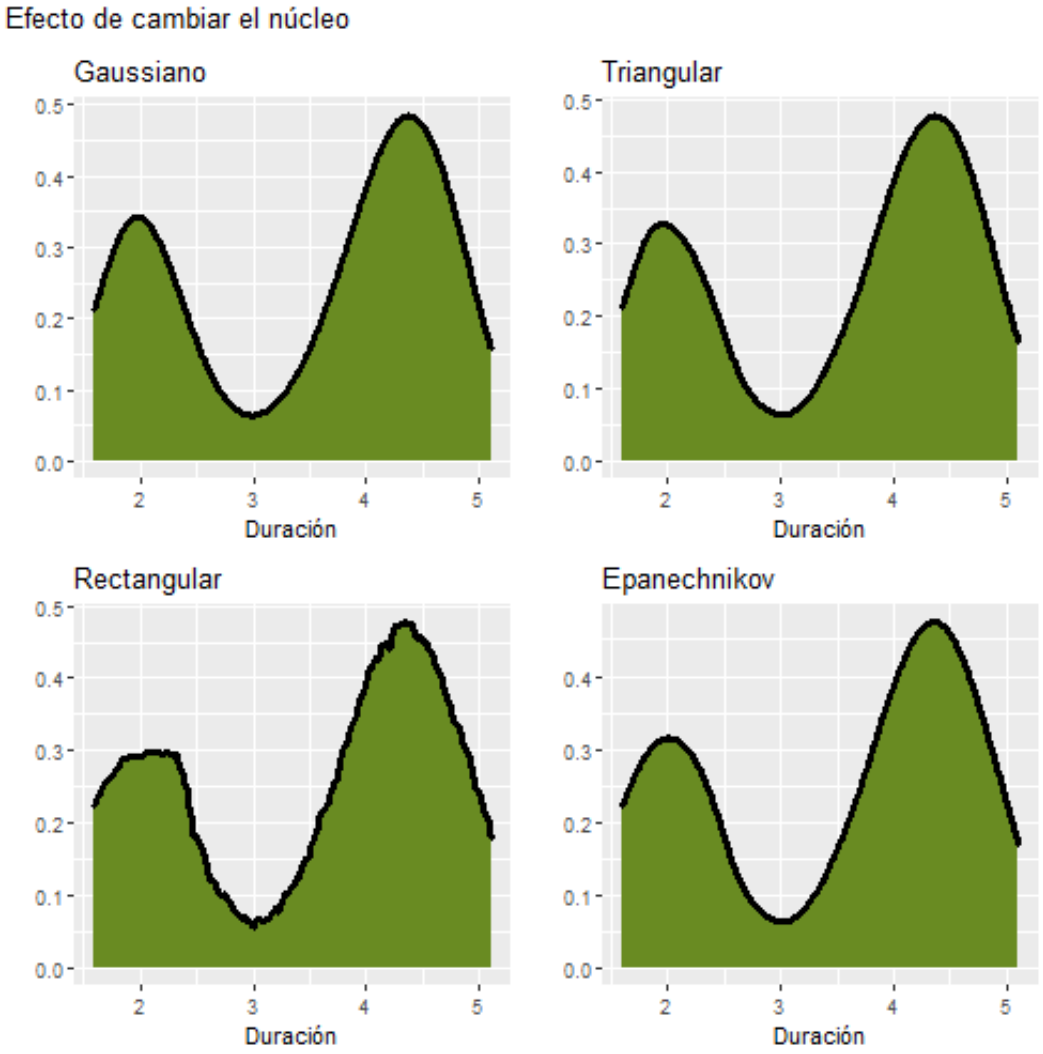
```
graf1 <- ggplot(faithful) +
  geom_point(aes(x = eruptions, y = waiting))

graf2 <- ggplot(faithful) +
  geom_density(aes(x = eruptions),
               bw = 0.15,
               fill = 'olivedrab4',
               size = 1.2)

graf1 / graf2
```

Representación con `geom_density`

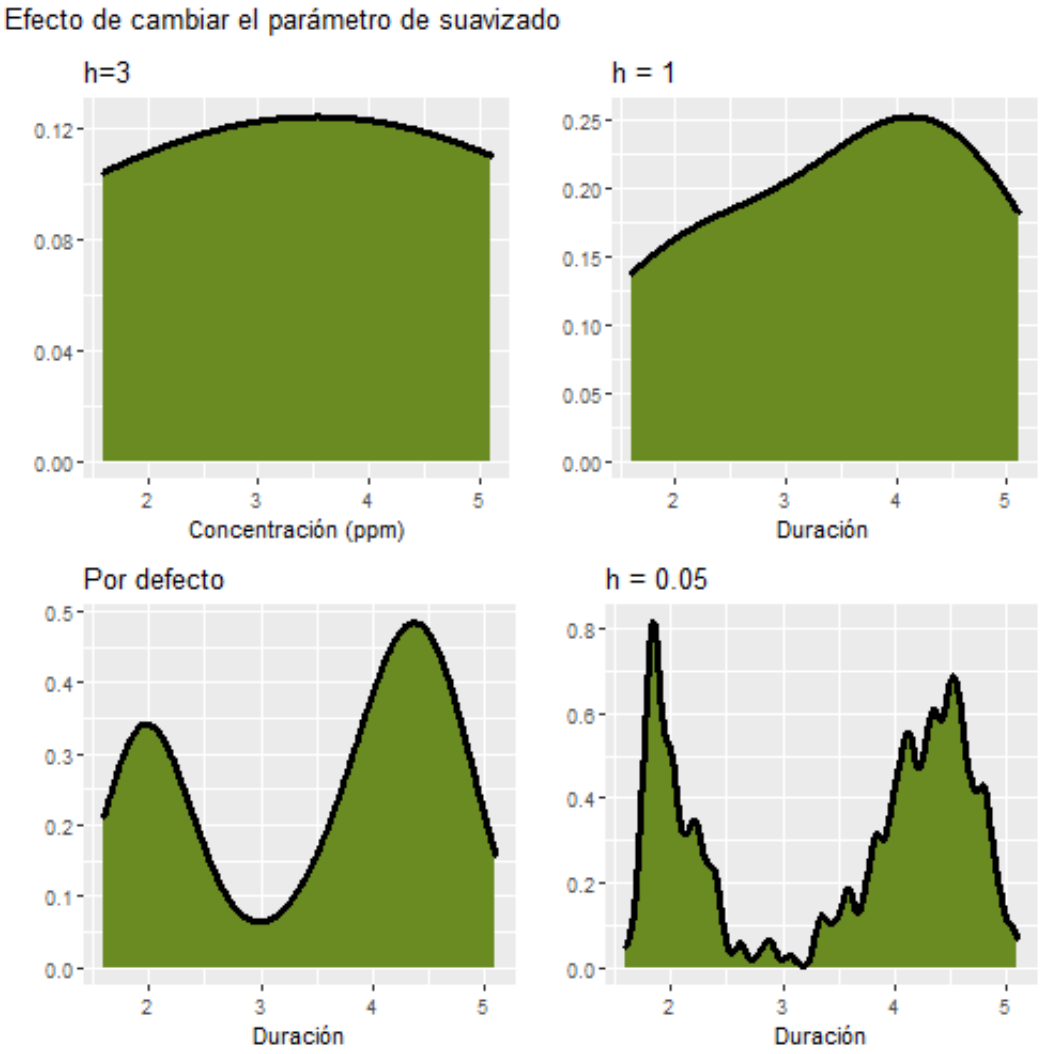
Fichero `faithful`: duración de las erupciones y del tiempo hasta la siguiente erupción del geyser *Old Faithful*



Efecto del núcleo



Efecto del parámetro de suavizado



Estimadores del núcleo como convolución

Definimos $U = Y + Z$, donde

- Y se distribuye de acuerdo con la distribución empírica F_n
- Z tiene densidad $K_h(x) = h^{-1}K(x/h)$
- Y y Z son independientes

Entonces, la densidad de la v.a. U es el estimador del núcleo $\hat{f}$

Un algoritmo para obtener vaiid con densidad $\hat{f}$,

1. Sortear con probabilidad $1/n$ entre $X_1, \ldots, X_n$. Supongamos que el resultado del sorteo es X^*.
2. Usar algún método de simulación estándar para obtener Z con distribución K_h.
3. Calcular $U = X^* + Z$.

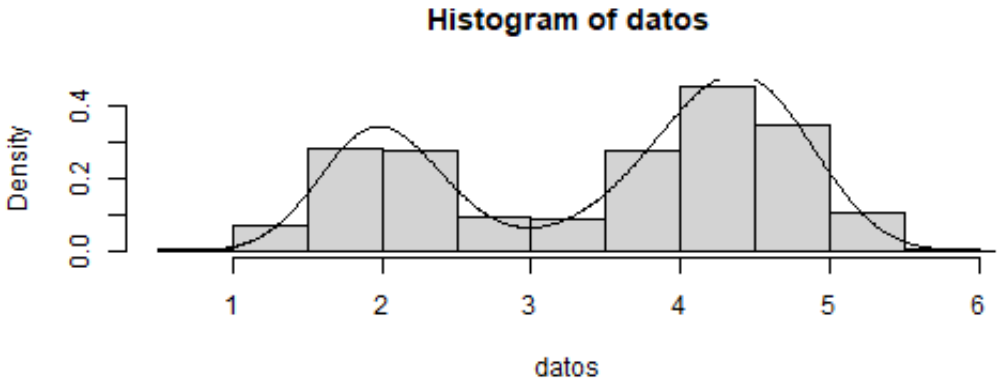
Ejercicio

Escribe un programa que genere realizaciones de una v.a. con densidad $\hat{f}$, donde $\hat{f}$ es el estimador del núcleo con núcleo gaussiano y parámetro de suavizado h. Aplícalo al ejemplo con los datos del geyser *Old Faithful*.

```
rnucleo <- function(n, muestra, h){
  # genera n observaciones de una distribución correspond
  # del núcleo (calculado con 'muestra') con núcleo gauss
  y = sample(muestra, n, rep = TRUE) + rnorm(n, sd = h)
  return(y)
}


muestra <- faithful$eruptions
estimador_nucleo <- density(muestra)
h <- estimador_nucleo$bw

hist(muestra)

datos <- rnucleo(10000, muestra, h)
hist(datos, freq = FALSE)
lines(estimador_nucleo$x, estimador_nucleo$y)
```

Ejercicio


Histogram of muestra
Frequency
0
40
2
3
4
5
muestra


Histogram of datos
Density
0.0
0.2
0.4
1
2
3
4
5
6
datos

Error cuadrático medio integrado

Para un valor fijo de x:

- El sesgo de $\hat{f}\,(x)$ es $\mathrm{E}[\hat{f}\,(x)] - f(x)$.
- La varianza es de $\hat{f}\,(x)$ es $\mathrm{E}[(\hat{f}\,(x) - \mathrm{E}(\hat{f}\,(x))^2]$.

El **error cuadrático medio** $\mathrm{E}[(\hat{f}\,(x) - f(x))^2]$ verifica:

$$\mathrm{ECM}(x) = \mathrm{Sesgo}^2[\hat{f}\,(x)] + \mathrm{Var}[\hat{f}\,(x)]$$

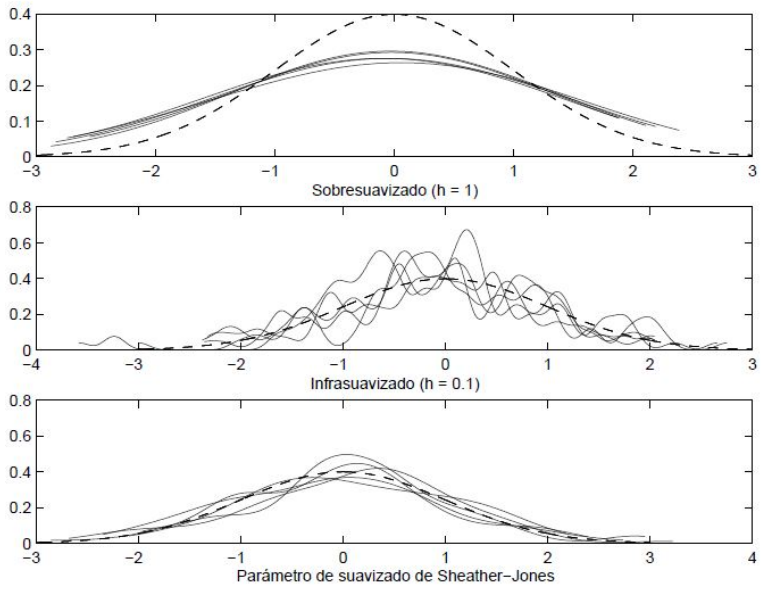
Queremos estimar la función de densidad f globalmente. Por ello, consideramos el error cuadrático medio **integrado**:

$$\mathrm{ECMI}(\hat{f}\,) = \int \mathrm{ECM}(x)dx = \int \mathrm{Sesgo}^2(x)dx + \int \mathrm{Var}[\hat{f}\,(x$$

Depende sobre todo del parámetro de suavizado h, del tamaño muestral n y de la suavidad de f

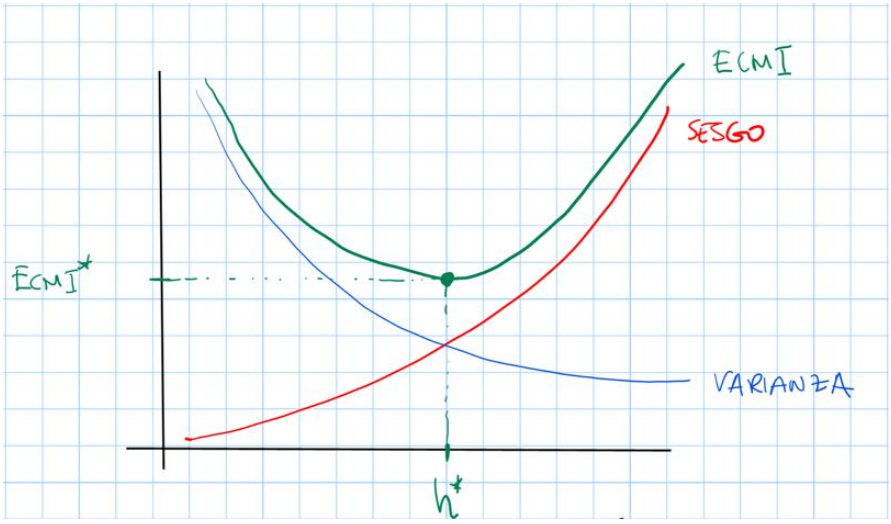
Error cuadrático medio integrado

Se trata de seleccionar $h = h_n$ de forma que se equilibren el sesgo y la varianza.

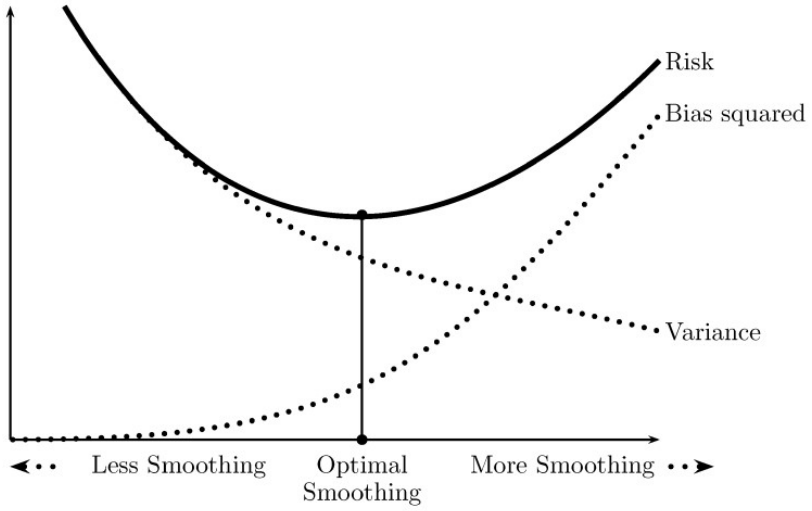


Error cuadrático medio integrado

Se trata de seleccionar $h = h_n$ de forma que se equilibren el sesgo y la varianza.



Error cuadrático medio integrado



(Fuente: Wasserman, L. (2013). *All of statistics: a concise course in statistical inference.* Springer.)

Otros criterios

Distancias L_p entre funciones:
$\|f-g\|_p = \left(\int |f(x)-g(x)|^p dx\right)^{1/p}$

El ECMI es el valor esperado de la distancia L_2 entre f y $\hat{f}$ al cuadrado:

$$\mathrm{ECMI}(\hat{f}) = \mathrm{E}(\|f-\hat{f}\|_2^2)$$

Dado que no todas las funciones de densidad verifican $\int f(x)^2 dx < \infty$, algunos autores proponen usar criterios L_1, como el error **absoluto** medio integrado:

$$\mathrm{EAMI}(\hat{f}) = \mathrm{E}(\|f-\hat{f}\|_1) = \mathrm{E}\left[\int |\hat{f}(x)-f(x)|dx\right]$$

La teoría para este tipo de criterios es más complicada debido a que el valor absoluto no es derivable

Aproximaciones del sesgo y de la varianza

Supuestos:

- K es una función simétrica, con $\int K(u)du = 1$, $\int uK(u)du = 0$, $\sigma_K^2 = \int u^2K(u)du < \infty$ y $d_k = \|K\|_2^2 = \int K(u)^2du < \infty$
- f es derivable dos veces con derivada continua

Aproximación del **sesgo**:

$$\text{Sesgo}^2[\hat{f}\,(x)] \approx \frac{h^4}{4}(f''(x))^2\sigma_K^4, \;\; \text{si} \;\; h \approx 0.$$

Aproximación de la **varianza**:

$$\text{Var}[\hat{f}\,(x)] = \frac{1}{nh}f(x)d_K, \;\; \text{si} \;\; nh \text{ grande}$$

ECMI aproximado:

$$\text{ECMI}(\hat{f}\,) \approx \frac{h^4}{4}\sigma_K^4\int f''(x)^2dx + \frac{d_K}{nh} = \frac{h^4}{4}\sigma_K^4\|f''\|_2^2 + \frac{\|}{}$$

Observaciones

- El término principal del sesgo aumenta si h es grande
- El término principal del sesgo también aumenta con la curvatura de f
- La varianza aumenta si h es pequeño y disminuye si el valor de nh es grande
- El término principal de la varianza no depende de f
- Un estimador consistente en el sentido L_2 requiere $h_n \to 0$ y $nh_n \to \infty$: bajo estas condiciones y condiciones de regularidad sobre f,

$$\lim_{n\to\infty} \mathrm{ECMI}(\hat{f}\,) = 0$$

Parámetro de suavizado óptimo

Una posible estrategia para seleccionar h es elegir el valor para el que se minimiza el ECMI aproximado:

$$h^* = \left(\frac{\|K\|_2^2}{\sigma_K^4 \|f''\|_2^2} \right)^{1/5} n^{-1/5}$$

y sustituyendo este valor en la expresión aproximada del ECMI tenemos

$$\text{ECMI}^* \approx \frac{5}{4} \sigma_K^{4/5} \|K\|_2^{8/5} \|f''\|_2^{2/5} n^{-4/5}$$

- La expresión de h^* sugiere que la velocidad con la que debe decrecer a cero h a medida que el tamaño muestral aumenta es $n^{-1/5}$, bastante lenta.
- El uso en la práctica de la fórmula es limitado puesto que h^* depende de $\|f''\|_2^2$, que es desconocida
- La tasa con la que ECMI va a cero es $n^{-4/5}$, más lenta que la tasa paramétrica habitual

Núcleo óptimo

En ECMI^*, la parte que depende del núcleo es $\phi(K) = \sigma_K^{4/5} \|K\|_2^{8/5}$ (invariante por cambios de escala en K)

Problema de cálculo de variaciones:

$$\min \|K\|_2^2 \text{ s.a. } K \geq 0, \int K(u)du = 1, \int uK(u)du = 0, \dots$$

La solución de este problema es el llamado *núcleo de Epanechnikov*:

$$K^*(u) = \frac{3}{4\sqrt{5}}\left(1 - \frac{t^2}{5}\right), \quad \text{si} \quad -\sqrt{5} \leq u \leq \sqrt{5}$$

La elección del núcleo no afecta mucho al ECMI: el cociente $\phi(K^*)/\phi(K)$ suelen estar por encima de 0.9 :

- Núcleo rectangular: 0.9295
- Núcleo gaussiano: 0.9512

En la práctica, se suele usar un núcleo gaussiano.

La moda muestral

En estadística elemental se suele definir la moda como el valor que más se repite. Esta definición es insatisfactoria para distribuciones continuas para las que no hay apenas empates.

Supongamos que la densidad f tiene una única **moda poblacional** θ tal que $f(\theta) = \max_{x\in\mathbb{R}} f(x)$.

Dado un estimador del núcleo $\hat{f}$ se define la **moda muestral** como $\hat{f}\,(\hat{\theta}) = \max_{x\in\mathbb{R}} \hat{f}\,(x)$.

Bajo ciertas condiciones de regularidad hay consistencia de la moda muestral para la poblacional:

Teorema. Sea $\hat{f}$ el estimador del núcleo obtenido a partir de una muestra de v.a.i.i.d. $X_1, \ldots, X_n$ de una distribución con función de densidad f. Se supone que f es uniformemente continua y la moda poblacional es única. Si $h_n \to 0$ y $nh_n^2 \to \infty$ entonces $\hat{\theta} \to_p \theta$.

Selección del parámetro de suavizado

Método plug-in suponiendo normalidad

- Sustituir $\|f''\|_2^2 = \int |f''(x)|^2 dx$ en la expresión de h^* por una estimación basada en el supuesto de que f es la densidad de una v.a. normal
- En el caso normal, $\|f''\|_2^2 = 3/(8\sqrt{\pi}\sigma^5)$.
- Se reemplaza este valor en h^* (núcleo K gaussiano):

$$h^* = \sigma(4/(3n))^{1/5} \approx \sigma(1.0456)n^{-1/5}$$

- Basta estimar σ para estimar h^*.
- Silverman propone usar $\hat{\sigma} = \min\{s, \hat{\sigma}_{ri}\}$, donde $\hat{\sigma}_{ri}$ es el rango intercuartílico (estandarizado para que converja a σ)

Selección del parámetro de suavizado

Método plug-in suponiendo normalidad

```
set.seed(100)
n <- 100
x <- rnorm(n)
bw.nrd(x = x)
```

```
## [1] 0.3982919
```

```
# El mismo resultado (salvo redondeo) que
ri <- diff(quantile(x, c(0.25, 0.75))) / diff(qnorm(c(0.2
1.0456 * n^(-1/5) * min(sd(x), ri)
```

```
## [1] 0.390266
```

```
# Una ligera variación (se multiplica por 0.9 en lugar de
bw.nrd0(x = x)
```

```
## [1] 0.3381724
```

Selección del parámetro de suavizado

Método plug-in no paramétrico

- Se fija un parámetro de suavizado preliminar g (usando por ejemplo la regla descrita en el apartado anterior) para obtener un estimador auxiliar $\hat{f}_g(x)$
- Se usa $\hat{f}_g$ para estimar $\widehat{\|f''\|_2^2} = \int \hat{f}_g''(x)^2 dx$.
- El método de Sheather y Jones es un refinamiento de esta idea (implementado en R: `bw.SJ`). Es uno de los métodos más recomendados.

```
bw.SJ(x = x)
```

```
## [1] 0.3663063
```

Selección del parámetro de suavizado

Validación cruzada

La muestra se divide en dos partes y se usa una de ellas para obtener información del procedimiento calculado con la otra. Este esquema se puede repetir muchas veces y se promedian los resultados.

$$\mathrm{ECMI}(\hat{f}\,) = \mathrm{E}\left[\int \hat{f}\,(x;h)^2 dx\right] - 2\mathrm{E}\left[\int \hat{f}\,(x;h) f(x) dx\right] +$$

El último término no depende de h, el objetivo es minimizar en h la estimación

$$C(h) = \int \hat{f}\,(x;h)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{\,(-i)}(X_i;h)$$

donde $\hat{f}_{\,(-i)}$ denota el estimador calculado con todas las observaciones a excepción de X_i.

Selección del parámetro de suavizado

Validación cruzada

Para determinar el parámetro de suavizado se resuelve numéricamente:

$$\hat{h} = \arg\min_{h>0} C(h) = \arg\min_{h>0} \left[\int \hat{f}\,(x;h)^2 dx - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{(-i)}(
$$

El problema es difícil ya que $C(h)$ puede tener muchos mínimos locales

Está implementado en `bw.ucv`:

```
bw.ucv(x = x)
```

```
## [1] 0.4224203
```

Cálculo con `density`

Los valores numéricos del estimador $\hat{f}(x)$ se calculan con `density()`

Cálculo de $\hat{f}(x)$ en cinco puntos equiespaciados en el intervalo [4,5]:

```
estimador_nucleo <- density(faithful$eruptions, n = 5, fr
estimador_nucleo$x # puntos donde se calcula la densidad
```

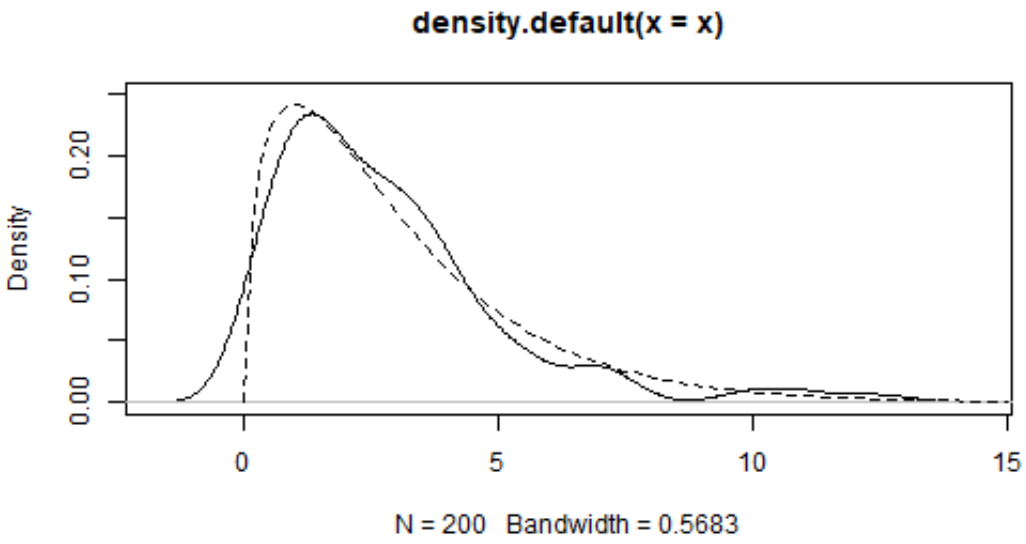
```
## [1] 4.00 4.25 4.50 4.75 5.00
```

```
estimador_nucleo$y # densidad estimada en esos puntos
```

```
## [1] 0.3853593 0.4717124 0.4701048 0.3667783 0.2143957
```

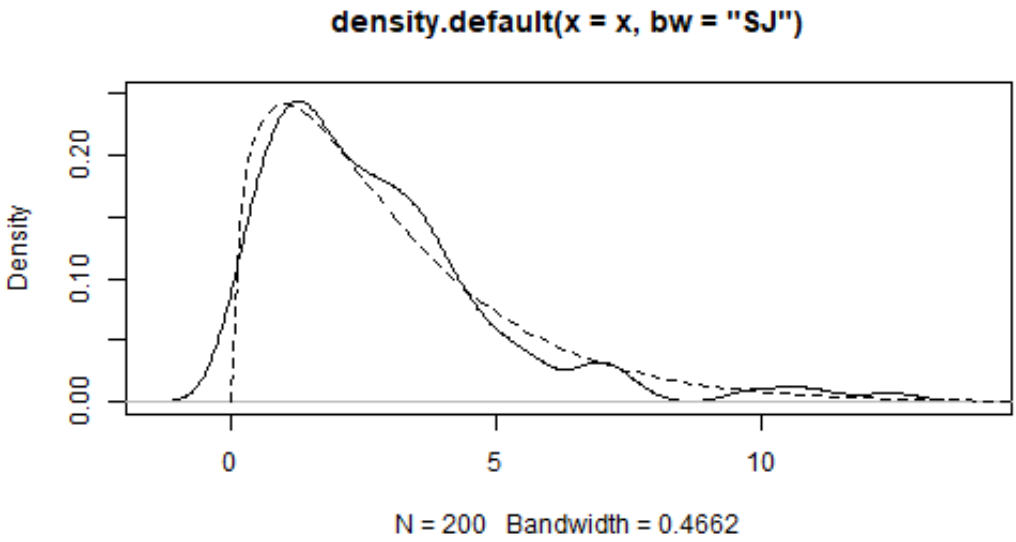
Cálculo con `density`

```
set.seed(100)
# Generamos n datos de distribución chi2 con 3 gl
n <- 200
x <- rchisq(n, 3)

# Opciones por defecto (bw = 'bw.nrd0' y núcleo gaussianc
nucleo <- density(x)
plot(nucleo, ylim = c(0,0.25))
curve(dchisq(x,3), from=0, to=15, lty=2, add=TRUE)
```



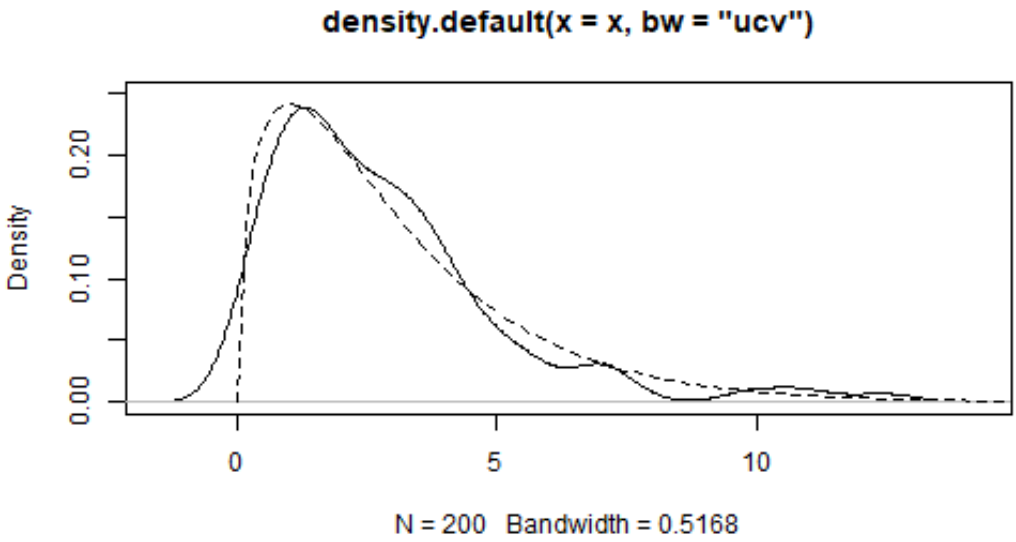
Cálculo con `density`

```
# Parámetro de suavizado de Sheather-Jones
nucleo_SJ <- density(x, bw = "SJ")
plot(nucleo_SJ, ylim = c(0,0.25))
curve(dchisq(x,3), from=0, to=15, lty=2, add=TRUE)
```



Cálculo con `density`

```
# Parámetro de suavizado por validación cruzada
nucleo_vc <- density(x, bw = "ucv")
plot(nucleo_vc, ylim = c(0,0.25))
curve(dchisq(x,3), from=0, to=15, lty=2, add=TRUE)
```



Estimación de densidades multivariantes

En dimensión d,

$$\hat{f}\,(x) = \frac{1}{n|H|}\sum_{i=1}^{n} \tilde{K}(H^{-1}(x - X_i)), \quad x = (x_1, \ldots, x_d) \in \mathbb{R}$$

donde $\tilde{K}$ es una densidad multivariante, Σ es una matriz $d \times d$ definida positiva, $H = \Sigma^{1/2}$ y $|H|$ es el determinante de H.

Simplificaciones:

- El núcleo $\tilde{K}$ es producto de núcleos unidimensionales idénticos:

$$\tilde{K}(x_1, \ldots, x_d) = K(x_1) \cdots K(x_d)$$

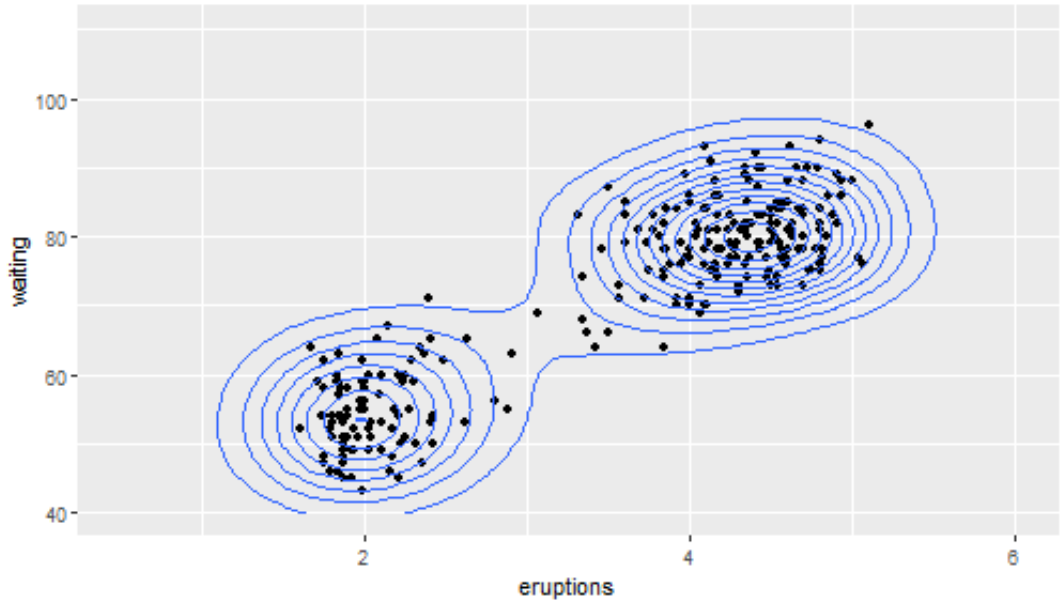
- La matriz H es diagonal y además el parámetro de suavizado es el mismo para todas las variables (es decir, $H = h\mathbb{I}_d$)

$$\hat{f}\,(x) = \frac{1}{nh^d}\sum_{i=1}^{n}\prod_{j=1}^{d} K\left(\frac{x_j - X_{i,j}}{h}\right), \quad x = (x_1, \ldots, x_d) \in$$

Estimación de densidades multivariantes

Para visualizar las curvas de nivel se usa `geom_density_2d` (producto de núcleos gaussianos con el mismo parámetro de suavizado en cada coordenada, seleccionado mediante `bw.nrd`). Más detalles: aquí y aquí.

```
ggplot(faithful, aes(x = eruptions, y = waiting)) +
  geom_point() +
  xlim(0.5, 6) + ylim(40, 110) +
  geom_density_2d()
```



La maldición de la dimensionalidad

- A no ser que se disponga de tamaños muestrales enormes, en dimensiones altas es difícil encontrar datos en muchas zonas del espacio muestral.
- Esto hace que se deterioren las propiedades de los estimadores.
- Por ejemplo, bajo condiciones de regularidad poco exigentes, puede probarse la siguiente propiedad de la aproximación asintótica al ECMI óptimo en dimensión d:

$$\mathrm{ECMI}^* \approx O(n^{-4/(4+d)})$$

- Este resultado muestra que la convergencia a cero del ECMI se hace más lenta a medida que la dimension crece.